Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Multivariate Analysis 90 (2004) 1–18

Journal of Multivariate Analysis

http://www.elsevier.com/locate/jmva

Determining and analyzing differentially expressed genes from cDNA microarray experiments with complementary designs

Erin M. Conlon,[a] Patrick Eichenberger,[b] and Jun S. Liu[a,*]

[a] *Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA*
[b] *Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA*

Received 31 March 2003

Abstract

We present a hierarchical Bayes model for determining genes in the sporulation pathway of *Bacillus subtilis* (*B. subtilis*) using two complementary replicated cDNA microarray experimental designs. The first design involves the mutation of a transcription factor, Sigma factor E (σ^E), and the second is an overexpression of this factor. We first normalize the microarray data using a rank invariant method. Genes found to be overexpressed in both experimental designs are further examined experimentally to determine their role in the sporulation pathway. Through statistical and experimental methods we found 181 genes that had not been previously described as controlled by σ^E. We identify the chromosome locations of clusters of σ^E-controlled genes using a nearest neighbor scan-statistic, and determine *B. subtilis* functional categories that are over-represented in subsets of expressed genes.

AMS 2000 subject classifications: 62H10; 62H12; 92D99

Keywords: Bayesian models; cDNA microarrays; Functional category; Gene location cluster; MCMC; Scan statistics

1. Introduction

Bacillus subtilis (*B. subtilis*) has been studied with much interest recently due to its close relation to *Bacillus anthracis*, the causative agent of anthrax, and its

*Corresponding author. Fax: +1-617-496-8057.
E-mail address: jliu@stat.harvard.edu (J.S. Liu).

0047-259X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2004.02.007

importance to public health concerns. *B. subtilis* is a bacterium that sporulates as a response to starvation, and in the sporulated state can survive in extreme environmental conditions. A main goal of this study is to determine genes in the sporulation pathway, specifically the σ^E pathway. σ^E is a transcription factor that controls a group of sporulation genes but to date only a fraction of genes under its control have been identified. Two complementary replicated cDNA microarray experimental setups have been implemented to show the same genes under σ^E's control. In the first set of experiments, referred to as the "mutant" design, sporulating cells that contain a null mutation in the gene for σ^E were compared to sporulating cells that are wild type for σ^E. In the second set of experiments, called the "induction" design, σ^E is overexpressed in response to a specific inducer, i.e. cells that had been treated with the inducer were compared to untreated cells.

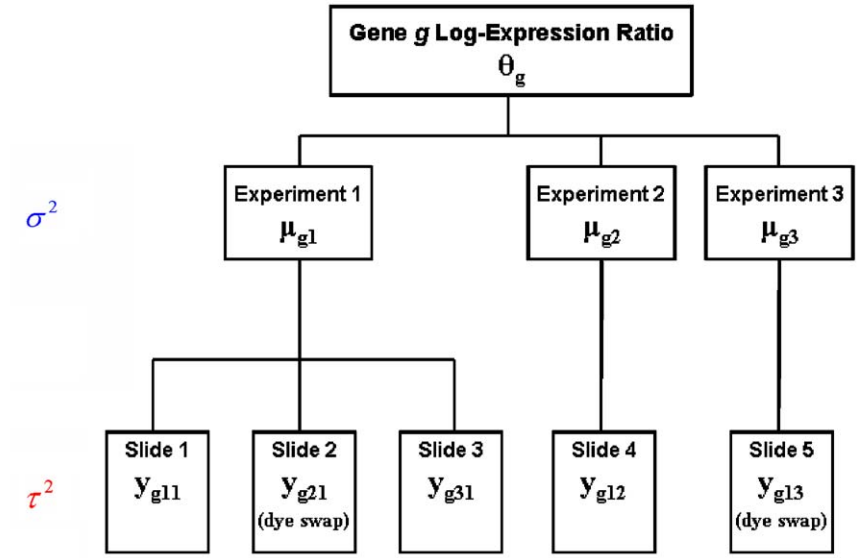
Within each design, a rank invariant method is first used to normalize individual slides. Then, a hierarchical Bayes model is employed to combine data across replicated slides to determine overexpressed genes. The analysis results of the two designs are combined to produce a consolidated set of 225 overexpressed genes. In addition to statistical methods, further biological insight identifies 28 additional putative genes for a final set of 253 genes predicted to be under the control of σ^E. Of the 253 genes, 72 were previously known to be in the σ^E regulon. Of the remaining 181 genes, bioinformatics and experimental methods confirmed that these are newly identified σ^E-controlled genes.

We perform a global test for clustering of σ^E-controlled genes along the *B. subtilis* genome, and identify chromosome locations of clusters using a nearest neighbor scan-statistic. We also identify functional categories that are over-represented among subsets of genes with specific expression characteristics using GeneMerge software [2], which produces *p*-values based on the Hypergeometric distribution. The categories of interest are the overexpressed, highest and lowest variance, and non-differentially expressed genes, based on the posterior distributions of log-expression ratios.

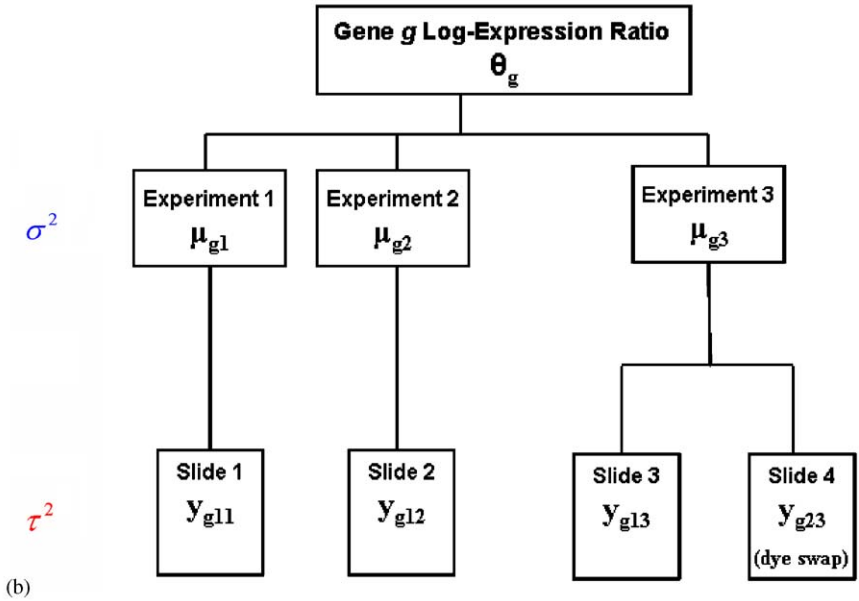
2. Methods

2.1. Mutant microarray experimental design

In the mutant design, five microarrays are produced from three independent experiments. Fig. (1a) displays a diagram of the experimental setup. In the first experiment, a culture of wild-type and mutant bacteria is separately grown and RNA is extracted from both cultures. The RNA from wild-type bacteria is used to generate fluorescently labeled cDNA tagged with Cy3 (green), and the RNA from mutant bacteria is used to generate cDNA labeled with Cy5 (red). These differentially tagged cDNA preparations are mixed and hybridized to slides one and three. Slide two is similar except that the Cy3 and Cy5 labels are reversed between the cultures (referred



(a)



(b)

Fig. 1. Experimental setup for (a) mutant design, and (b) induction design. σ^2 is the experimental variance, τ^2 is the slide variance.

to as a "dye swap"). Similarly, experiments 2 and 3 produce one slide each, with the last slide being a dye swap. We use wild-type/mutant ratios since up-regulation of these values will identify genes in the σ^E regulon.

The number of genes spotted on each of the five arrays ranges from 4268 to 4751. These numbers are larger than the 4106 genes in the *B. subtilis* genome since selected genes of interest were spotted multiple times on various arrays. Each of the slides contains low-quality spots that are flagged and removed from analysis. The percentage of genes detected on the arrays with adequate quality for analysis ranges from 35.5% to 81.4% (between 1687 and 3749 genes, respectively). We use mean foreground minus mean background intensity levels to calculate expression for both the Cy3 and Cy5 channels for each gene. Genes with negative intensity levels for both the Cy3 and Cy5 channels are removed from analysis. Genes with negative expression values (due to the subtraction of background values) in only one of the two channels are assigned a minimum positive intensity value of 10. This negative intensity value is typical in the σ^E mutant channel, in which many genes have close-to-background expression level.

2.2. Induction microarray experimental design

In the induction design, four microarrays are produced from three independent experiments (Fig. 1b). The first two experiments produce one slide each, in which cDNA from the wild-type bacteria culture is labeled with Cy5 and cDNA from the induced bacteria culture is labeled with Cy3. The third experiment produces two slides, with the last slide being a dye swap. We use induction/wild-type ratios since up-regulation of these values will show genes under the control of σ^E. Calculation of intensity values is similar to the mutant design.

The number of genes spotted on the four arrays ranges from 4608 to 4751. Selected genes were also spotted multiple times on the arrays, similar to the mutant design. The percentage of genes detected on the arrays ranges from 33.0% to 47.4% (between 1519 and 2183 genes). Note that these percentages are much lower than for the mutant design. Given that signal intensities (or at least the foreground/background ratios) were usually lower in the induction experiments, more features had to be flagged and removed from the analysis in the induction design than the mutant design.

2.3. Rank invariant normalization

Popular methods of normalization include the use of "housekeeping" genes, which are genes that are thought to be non-differentially expressed in an experimental condition. However, it has been shown that these genes show some natural variability and are thus not always well suited for normalization purposes [14]. Here, we use the rank invariant method of [12] to normalize each slide to remove systematic effects not related to gene expression. In this procedure, we first calculate the ranks of each gene among the Cy3 and Cy5 channels, respectively. Genes with Cy3 and Cy5 ranks within $2\%G$ and with average of Cy3 and Cy5 ranks not among the 25 highest or lowest ranks are included in the initial set of genes S_0. Here G is the total number of genes on the slide. This process is repeated, i.e., the set S_i contains

genes with ranks of Cy3 and Cy5 within $(2\%|S_{i-1}|)$, until the rank invariant set does not change. The value of $2\%G$ and 25 are based on recommendations by Tseng et al. [14], for similar sized data sets.

As recommended in [5], we compute for each gene the "M" and "A" values, where $M = \log_{10}$-ratio of expression (y-axis) and $A =$ average $\log_{10}$-intensity over the Cy3 and Cy5 channels (x-axis). A locally weighted scatterplot smoother "Lowess", a built-in function of the statistical language R [8], is used to fit a normalization curve $\hat{M}$ through the rank invariant genes, where $\hat{M} = \hat{f}(A)$. The normalized log-ratios are defined as $\tilde{M} = M - \hat{M}$. After within-slide normalization, we determine whether further normalization due to scale is needed by comparing the spread of the distributions of the slides within each of the mutant and induction designs.

2.4. Hierarchical Bayes model

We combine the data across normalized slides using the hierarchical Bayes model in [14]. This model takes into account variation due to slides and experiments, and allows missing gene expression data and genes that are multiply spotted on some arrays and not others. More specifically, the model assumes that:

$$y_{gse} \mid \mu_{ge} \sim N(\mu_{ge}, \tau_g^2), \quad g = 1 \ldots G;\ e = 1 \ldots . E;\ \ s = 1 \ldots S_e,$$
$$\mu_{ge} \sim N(\theta_g, \sigma_g^2), \quad g = 1 \ldots G;\ e = 1 \ldots E.$$

Here, y_{gse} is the observed log-ratio of expression for gene g, slide s and experiment e, and μ_{ge} is the mean over slide within experiment. The θ_g is the true log-ratio of expression for each gene and is the parameter of interest, τ_g^2 is the slide effect variance, σ_g^2 is the experiment effect variance, and S_e is the number of slides in experiment e.

It is difficult to validate the model assumption (normality) since each gene has its own set of mean and variance parameters and no more than five observations. We present here a novel method to test for normality. For each gene, we choose one observed measurement (if there are multiple) for each experiment (see Fig. 1b). Then, we compute the gene's normalized maximum deviance (NMD):

$$NMD_g = \frac{\max_i \mid x_{gi} - \bar{x}_{g.} \mid}{\sqrt{\frac{1}{E-1} \sum_{i=1}^{E} (x_{gi} - \bar{x}_{g.})^2}}.$$

Since the NMD is independent of the mean and variance, its null distribution can be simulated. We calculated NMD_g for both the induction experiment data (Fig. 1b) and simulated Normal(0,1) data and performed a Kolmogorov–Smirnov test to compare the two distributions. We obtained a p-value of 0.8605, indicating that the normality assumption is acceptable for our experiments.

As prior specifications, we assume that $p(\theta_g) \propto 1$; $\tau_g^2 \sim k\tilde{\tau}^2/\chi_k^2$; and $\sigma_g^2 \sim h\tilde{\sigma}^2/\chi_h^2$. The scale parameter representing the between-slide variation $\tilde{\tau}_g^2$ is derived from

the data:

$$\tilde{\tau}^2 = \frac{1}{G(\sum S_e - 1)} \sum_{g=1}^{G} \sum_{e=1}^{E} \sum_{s=1}^{S_e} (y_{gse} - y_{g\cdot e})^2,$$

where $y_{g.e}$ is the average log-expression ratio over the slides within an experiment:

$$y_{g\cdot e} = \frac{1}{S_e} \sum_{s=1}^{S_e} y_{gse}.$$

Similarly, the between-experiment variation is given as

$$\tilde{\sigma}^2 = \frac{1}{G(E-1)} \sum_{g=1}^{G} \sum_{e=1}^{E} (y_{g\cdot e} - y_{g\cdot\cdot})^2,$$

where $y_{g..}$ is the average log-expression ratio over both slides and experiments. We used three degrees of freedom for both τ_g^2 and σ_g^2, i.e. $h=k=3$. The posterior computation is accomplished by a Gibbs sampler [10], which cycles through the full conditional distributions:

$$p(\mu_{ge} \mid y_{gse}, \theta_g, \tau_g^2, \sigma_g^2) \propto N\left(\frac{S_e y_{g\cdot e}\sigma_g^2 + \tau_g^2\theta_g}{S_e\sigma_g^2 + \tau_g^2}, \frac{\tau_g^2\sigma_g^2}{S_e\sigma_g^2 + \sigma_g^2}\right),$$

$$p(\theta_g \mid y_{gse}, \mu_{ge}, \tau_g^2, \sigma_g^2) \propto N(\mu_{g\cdot}, \sigma_g^2/E),$$

$$p(\sigma_g^2 \mid y_{gse}, \mu_{ge}, \theta_g, \tau_g^2) \propto \left\{\sum_{e=1}^{E} (\mu_{ge} - \mu_{g\cdot})^2 + h\tilde{\sigma}_g^2\right\} \Big/ \chi^2_{E+h-1},$$

$$p(\tau_g^2 \mid y_{gse}, \mu_{ge}, \theta_g, \sigma_g^2) \propto \left\{\sum_{e=1}^{E} \sum_{s=1}^{S_e} (y_{gse} - \mu_{ge})^2 + k\tilde{\tau}_g^2\right\} \Big/ \chi^2_{S_1+\cdots+S_E+k}.$$

We implemented the foregoing procedure using the *R* statistical language [8], with subroutines provided by Tseng et al. [14]. The number of Gibbs sampling iterations was chosen as 2000 for our problems, which is more than adequate. The posterior distribution of θ_g, in particular $P(\theta g > 0|y)$, is used for the inference of appreciably overexpressed genes.

2.5. Combining two experimental designs

We define a gene's *score* as the posterior probability that its true log-expression ratio is greater than zero, using the model in Section 2.4. Overexpressed genes in each experimental design are those genes with scores greater than a threshold, which is selected by comparison to the scores associated with genes that had been previously identified as controlled by σ^{E}. The threshold value for the mutant design was 0.95, i.e. genes with 95% or greater posterior probability of a positive log-expression ratio are defined as overexpressed in the mutant design. The induction experiments produced lower expression ratios resulting in a less stringent threshold value of 0.85. Differentially expressed genes are those with the indicator

variable $D_g = 1$, where

$$D_g = \begin{cases} 1 & \text{if } (S_{gm} > 0.95) \cap (S_{gi} > 0.85), \\ 0 & \text{otherwise,} \end{cases}$$

with S_{gm} and S_{gi} being the scores for gene g in the mutant and induction designs, respectively.

In the case of no prior knowledge of genes under the control of a transcription factor, the following procedure may be used to determine a set of target genes. First, a stricter cutoff would be used as a starting point, and a small number of genes would be identified as differentially expressed. These genes would be experimentally verified to identify a short consensus sequence (called a binding site) in the regulatory region. It is also possible to use computational means to discover these consensus sequences [3]. One would then explore the genes left out, performing pattern matching of the consensus sequence in the regulatory sequence of genes below the first cutoff. These genes would then be experimentally targeted and verified. This process would iterate until there were a larger set of genes that the experimentalist was confident were under the control of the transcription factor.

2.6. Clusters of σ^E-controlled genes

2.6.1. Global clustering test

The genome of *B. subtilis* is comprised of one circular chromosome containing 4106 genes, among which 252 have been determined by our analysis as controlled by σ^E and annotated in the *B. subtilis* genome (see Sections 3.3 and 3.4). To test whether these 252 genes are scattered in the genome uniformly, we conduct both the gap and the cluster analyses. For display purposes, we break the chromosome at the origin, which is located 410 bases upstream from the gene *dnaA*. Regarding the 4106 gene locations as equally spaced along a unit circle, we plot the histogram of the spaces between the neighboring σ^E-controlled genes (figure not shown). If these 252 genes were indeed uniformly distributed around the circle, the spacing distribution should follow a *Beta*(1, 251) distribution. We examine the quantiles of the *Beta*(1, 251) distribution to determine how well the data fits the null distribution (see Results).

2.6.2. Cluster locations

To identify genomic locations of clusters of σ^E-controlled genes, we define for each gene g the statistic SNN_g (**S**igma E-controlled **N**earest **N**eighbors) as the number of σ^E-controlled genes within the 50 genes centering at gene g, i.e.,

$$SNN_g = \sum_{i=g-25}^{g+25} I_i, \quad \text{where } I_i = \begin{cases} 1 & \text{if gene } i \text{ is } \sigma^E\text{-controlled,} \\ 0 & \text{otherwise.} \end{cases}$$

SNN_g is a "scan statistic" on the circle, i.e. the maximum number of points in a moving window of fixed length over a region. The neighborhood size 50 was chosen from a priori information of cluster sizes based on comparative genomics. This value can be tailored by the user for specific study questions. Exact tail probabilities for the

scan statistic on the line have been calculated for small numbers of events (<25, see [11]), assuming an underlying uniform distribution of events on the interval and a continuous window length. Asymptotic results for these exact probabilities have been derived for the line and circle [4]; however, the results are impractical to compute and are not accurate enough to be used for testing purposes [7]. More recent accurate approximations of tail probabilities of the scan statistic on the line have been introduced that are computationally feasible (see [7] and references therein). However, these methods have been extended neither to the circle nor to the case of discrete grids, as we have here for the circular bacterial genome. We therefore determine significance levels for the *SNN* statistic through a Monte Carlo (MC) simulation. We first label randomly 252 genes out of the 4106 as being controlled by σ^E, and then calculate the maximum of the *SNN* statistic over all the 4106 genes. This process is repeated 1000 times, resulting in 1000 maximum values. The 95% and 99% quantiles of these 1000 values are shown in Fig. 5, together with the plot of the observed *SNN* statistics for each gene versus its chromosomal location.

2.7. Functional categories

The functional classification tree of the genes of *B. subtilis* contains six functional categories, including: (1) cell envelope and cellular processes, (2) intermediary metabolism, (3) information pathways, (4) other functions, (5) similar to unknown proteins, and (6) unknown. Within these six categories are 46 subcategories. We identify functional categories that are over-represented in subsets of expressed genes using the GeneMerge software [2], which computes the *p*-value of over-represented categories based on the hypergeometric distribution. For example, if the total number of genes detected on a microarray slide is N, among which M genes belong to category K. Suppose a subgroup of n genes possess certain expression characteristics such as overexpression. If category K is not related to the genes that are affected in the experiment, we expect that the number k of the overexpressed genes that belong to category K follows the hypergeometric distribution $Hyper(N,M,n)$. Since many categories are considered, the *p*-values need to be adjusted for multiple testing with Bonferroni corrections.

3. Results

3.1. Mutant design

We display the pre-normalized M–A plots for the genes detected on the arrays for each of the five slides (Fig. 2). In all five slides, the distributions of the log-ratios show higher expression in the Cy3 than the Cy5 dye. This is consistent with the better labeling efficiency of Cy3 that has been documented in many experiments [5]. We remove this effect through rank invariant normalization. We find between 3% and 6% of the total genes (between 47 and 218 genes) are rank invariant for each of the

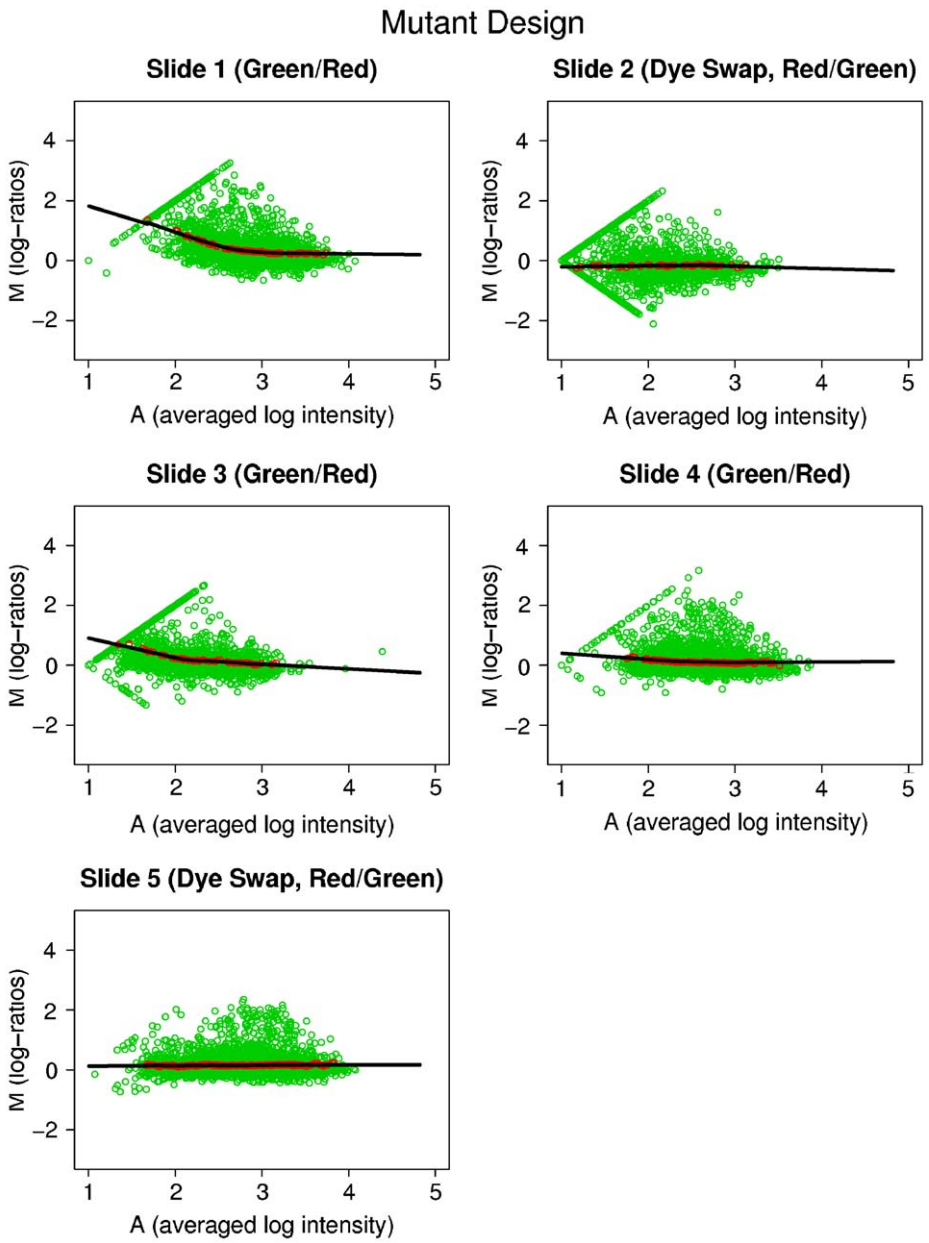


Fig. 2. Pre-normalized $\log_{10}$-ratios of expression, the rank invariant genes, and the Lowess curve fit through the rank invariant genes for the mutant experimental design.

slides, and we fit a Lowess normalization curve through these genes for each slide (Fig. 2). Note that the apparent straight lines of the M–A plots are due to the minimum positive intensity value assigned to the genes with negative intensity in one

channel. The boxplots of each of the five slides are examined to determine whether further normalization is needed due to scale (Fig. 3). The boxplots show that slide 2 has a larger spread in its distribution than the other slides. Slide 2 also had the lowest percent of genes of sufficient quality for analysis (35.5%). However, the five distributions are adequately similar that normalizing across arrays may introduce more variance than not adjusting [15]. As a result we combine the normalized data as it is, using the hierarchical Bayes model of Section 2.4. Using the 0.95 score threshold, i.e. requiring overexpressed genes to have 95% or higher posterior probabilities of positive log-expression ratios, we find 372 genes that are overexpressed in the mutant design.

3.2. Induction design

Implementation of the analysis methods for the induction design is similar to that for the mutant design. We examine the pre-normalized M–A plots for each of the four slides (figure not shown). In all four slides, the distributions of the log-ratios again show the effect of higher expression in the Cy3 than the Cy5 dye, which is removed through rank invariant normalization. We find between 5% and 8% of the genes are rank invariant (between 81 and 143 genes) for each of the slides, which is a slightly higher percentage than for the mutant design. The rank invariant genes and

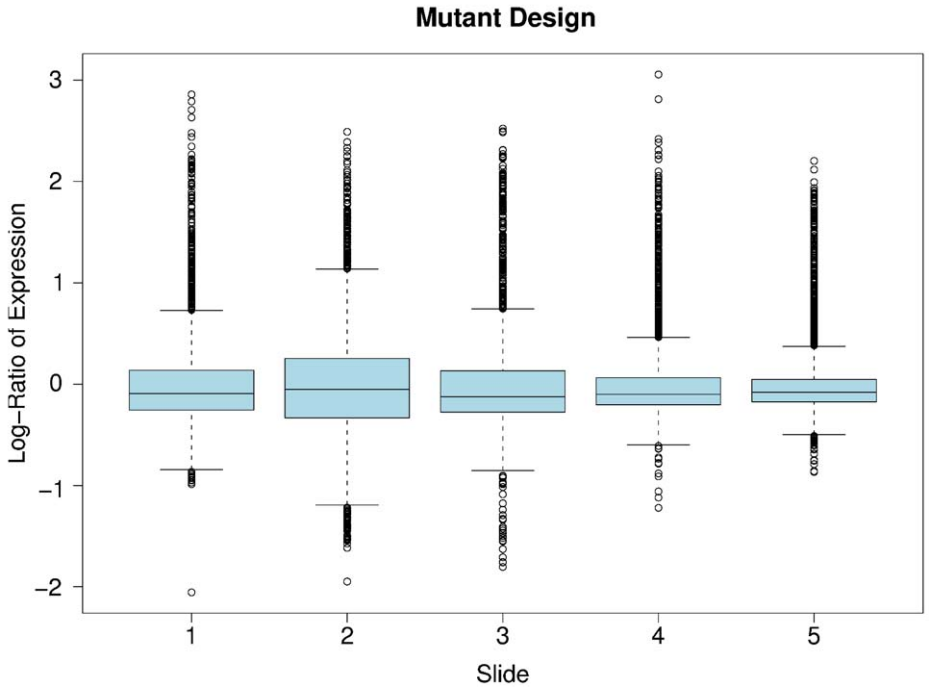


Fig. 3. Boxplots of distributions of $\log_{10}$-ratios of expression after normalization for the mutant design, five wild-type/mutant micorarrays.

the Lowess normalization curves are also calculated for each slide. The boxplots for the four slides are similar, so that we do not normalize for scale (figure not shown). Using the 0.85 score threshold, i.e. defining overexpressed genes as those with 85% or higher posterior probability of positive log ratios of expression, we find 329 genes that are overexpressed in the induction design.

3.3. Overexpressed genes in both mutant and induction designs

We find $D_g = 1$ for 225 genes of *B. subtilis*, which are candidates for control by σ^E. Further biological knowledge of 28 additional genes increased the total putative genes to 253 [6], as follows. Of the 28 genes, 16 were previously known to be controlled by σ^E. Of the remaining 12, eight were above the score threshold in the mutant design, and had a predicted σ^E-binding site in the regulatory region, three were not identified in any experiments but belonged to operons (i.e. 2 or more genes that are next to each other and controlled by the same promoter) with other genes successfully identified, and one was not annotated in the genome.

3.3.1. Individual gene results, *spoVR* and *glyA*

We illustrate our method of combining the mutant and induction designs by using individual genes. For the mutant design, gene *spoVR* has posterior log-expression ratios ranging from 1.5 to 2.8 (Table 1). The histogram of the posterior distribution is shown in Fig. (4a), with posterior median of 2.17. *spoVR* has a score of 1.0, i.e. 100% of the posterior distribution is greater than zero, so that ($S_{gm} > 0.95$) and *spoVR* is classified as overexpressed in the mutant design. For the induction design, *spoVR* has posterior log-expression ratios ranging from 0.29 to 0.47, with posterior median of 0.39 (Table 1, Fig. 4b). The *spoVR* induction score is 0.996, so that ($S_{gi} > 0.85$) and *spoVR* is classified as overexpressed in the induction design. $D_g = 1$ for this gene, and thus *spoVR* is labeled differentially expressed and potentially under

Table 1
Normalized $\log_{10}$-expression ratios, posterior medians and scores for genes *spoVR* and *glyA*

	Mutant design		Induction design		
	Normalized $\log_{10}$-ratio *spoVR*	Normalized $\log_{10}$-ratio *glyA*		Normalized $\log_{10}$-ratio *spoVR*	Normalized $\log_{10}$-ratio *glyA*
Slide 1	2.22	−0.18	Slide 1	0.47	−0.04
Slide 2	1.64	0.47	Slide 2	0.35	0.05
Slide 3	2.52	−0.01	Slide 3	0.29	−0.04
Slide 4	2.81	0.18	Slide 4	0.38	−0.01
Slide 5	1.46	0.17			
Posterior median	2.17	0.16	Posterior median	0.39	−0.03
Score, S_{gm}	100%	77%	Score, S_{gi}	99.6%	38%

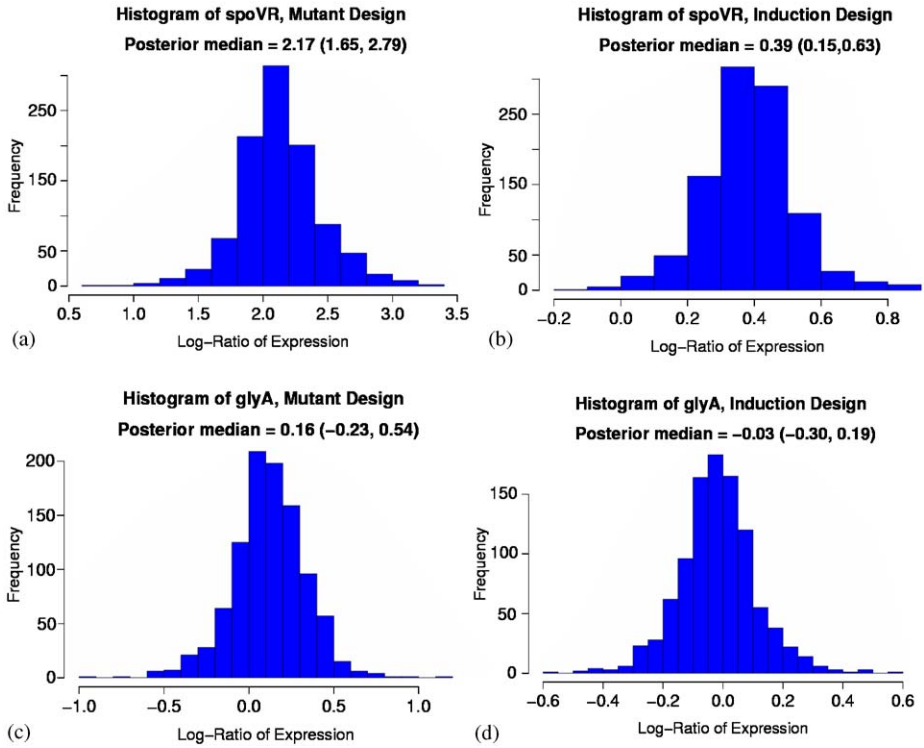


Fig. 4. Histograms of posterior distributions of two sporulation genes. *spoVR* (a) mutant design, (b) induction design, *glyA* (c) mutant design, (d) induction design. *spoVR* is differentially expressed, *glyA* is non-differentially expressed.

the control of σ^E, which is consistent with a previous report identifying this gene as a σ^E-controlled gene [1].

In contrast, the histogram of the posterior log-expression ratios of gene *glyA* for the mutant design is shown in Fig. (4c). This gene has posterior log-ratios ranging from -0.18 to 0.47, with median 0.16 (Table 1). The *glyA* mutant score is 0.77, so that ($S_{gm} < 0.95$) and *glyA* is classified as non-differentially expressed in the mutant design. For the induction design, *glyA* has posterior log-ratios ranging from -0.10 to 0.05, with posterior median of -0.03 (Table 1, Fig. 4d). The *glyA* induction score is 0.38, so that ($S_{gi} < 0.85$) and *glyA* is classified as non-differentially expressed in the induction design. $D_g = 0$ for this gene and thus *glyA* is labeled not under the control of σ^E. Note the much lower posterior log-expression ratios for the induction versus mutant design for both *spoVR* and *glyA*.

3.4. Experimental follow-up

The 253 putative σ^E-controlled genes were further examined computationally and experimentally to confirm their inclusion in the σ^E regulon. Seventy-two of the 253

genes were identified previously to be under the control of σ^E. Of the 72, 16 were from the set of 28 additional genes based on biological knowledge. Of the remaining 181 genes, further computational methods identified putative σ^E-binding sites in the respective regulatory regions. Twelve of these 181 genes were from the set of 28 additional genes based on biological knowledge. Computer predictions were confirmed by mapping of the promoters for several of the newly identified σ^E-regulated genes. In addition, systematic gene inactivation revealed that many of these genes are critical for efficient sporulation. Finally, experiments using in-frame fusions of several of the genes to the coding sequence for the green fluorescent protein (GFP) indicated that subcellular localization of the proteins encoded by these genes is consistent with a role in the process of sporulation [6]. Thus, as judged by bioinformatics and experimental methods, the 181 genes are newly identified σ^E-controlled genes. Of the 253 σ^E-controlled genes, one gene, *ytrH*, is a newly discovered gene that is not annotated in the *B. subtilis* genome version R16.1 (http://genolist.pasteur.fr/SubtiList/), so that the set analyzed below is size 252.

3.5. Gene clusters

We examine the histogram of the number of non-σ^E-controlled genes located between the 252 σ^E-controlled genes (i.e. the "gaps"), and compare the gap distribution with the *Beta*(1, 251) quantiles. Of the 252 gaps, 12 have values greater than the 99% quantile, and two have values greater than the 99.99% quantile. This indicates larger than expected gaps between σ^E-controlled genes. For the lower tail, 111/252 (44%) of the gaps equal zero, i.e. the σ^E-controlled genes are adjacent on the chromosome. This is highly significant even after adjusting for the existence of operons (it is reasonable to assume that no more than 50% of the 252 σ^E-controlled genes are in operons; that the average operon size is 2.8 genes [17]). Even though 142 of the 252 genes (i.e. 56%) could be grouped into operons [6], we find several adjacent pairs of σ^E-controlled genes that are transcribed in opposite orientation and are thus not organized in operons. We do not explicitly adjust for operons in this analysis since the operon organization has not been systematically confirmed and is only predicted. These findings are indicative of global clustering. One gap value greater than the 99.99% quantile is due to differences between bacterial strains used to build the microarray and that used in the experiments, and does not provide additional evidence of clustering (see Discussion). We also performed a one-sample two-sided Kolmogorov–Smirnov test, comparing the gap distribution to a *Beta*(1, 251) distribution. The p-value is 2.2×10^{-16}, indicating a highly significant departure from the *Beta*(1, 251) distribution.

We plot SNN_g, i.e. the number of σ^E-controlled genes within 50 nearest neighbors, for the 4106 genes of *B. subtilis* versus chromosome location, and the 95% and 99% Monte Carlo quantiles (Fig. 5). Of the 4106 SNN_g statistics, there are two major peaks that are above the 99% MC quantile. The locations of the largest clusters are between genes 2426 and 2437, all of which have 16 σ^E-controlled genes in the 50

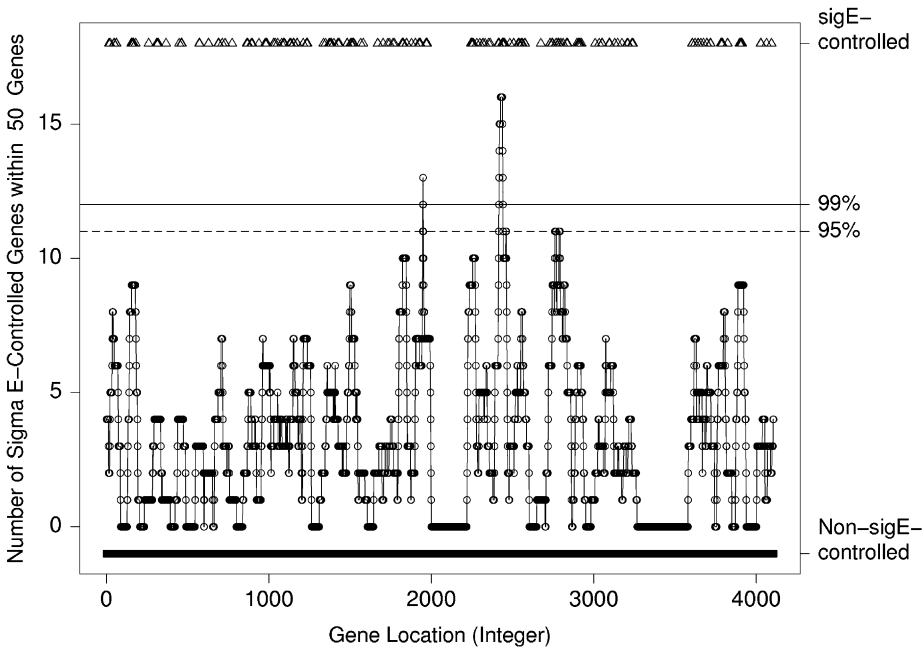


Fig. 5. SNN_g for each of the 4106 genes of the *B. subtilis* genome versus gene location. The 95% (dashed) and 99% (solid) lines correspond to Monte Carlo significance levels. Values above the 95% and 99% lines indicate significant clusters of σ^E-controlled genes.

nearest neighbors. Another large cluster at gene 1949 contains 13 σ^E-controlled genes within 50 nearest neighbors. Areas with no σ^E-controlled genes are also of interest, including a large segment of 312 genes from locations 3268 to 3579 with no σ^E-controlled genes. There is a second large segment of 218 genes between genes 2001 and 2218 with no σ^E-controlled genes. This segment is due to differences between bacterial strains used to build the microarray and that used in the experiments, and does not provide further evidence of clustering (see Discussion).

3.6. Gene function analysis

3.6.1. Overexpressed genes

We identify the functional subcategories that are over-represented for the 252 genes under the control of σ^E. The population for this statistical test is the set of 3746 genes detected on at least one array of the mutant or induction designs. The 252 genes represented 27 functional subcategories, of which two were over-represented with highly significant adjusted p-values (less than 0.001). The first is "sporulation", which is expected for this group of genes; the second is "metabolism of lipids". A third subcategory had a borderline significant adjusted p-value of 0.011, that of "cell

Table 2
Over-representation of functional categories of genes

Bacillus subtilis function	Population fraction	Study fraction	*p*-Value	Bonferroni-adjusted *p*-Value
Overexpressed genes				
1.1, Cell wall	0.075	34/252	0.00042	0.011
2.4, Metabolism of lipids	0.017	16/252	3.6e−06	0.0001
1.8, Sporulation	0.035	49/252	1.6e−25	4.1e−24
Lowest variance				
3.6, RNA modification	0.0059	11/433	1.0e−05	0.0004
1.4, Membrane bioenergetics	0.020	21/433	5.1e−05	0.0020
2.2, Metabolism of amino acids and related molecules	0.044	33/433	0.0011	0.042
1.1, Cell wall	0.075	49/433	0.0016	0.063
Highest variance				
6, unknown	0.137	107/473	8.8e−09	3.2e−07
Non-differentially expressed genes				
None				

The population fraction is the fraction of genes in the population with the associated function. The study fraction is the fraction of genes in the subset of interest with the associated function.

wall'' (Table 2). These findings suggest that genes involved in metabolism of lipids and cell wall functions are important in the σ^E sporulation pathway. Given that sporulation is initiated in response to starvation, it is not surprising that genes involved in metabolic functions are activated. Catabolism of lipids could provide part of the energy necessary to complete the process of sporulation under conditions of limited nutrient availability. Similarly, the presence of genes from the ''cell wall'' category is expected because the spore cortex, an elaborate structure that protects the spore against certain forms of adverse environmental conditions, is chemically and physically related to the cell wall. Spore cortex synthesis takes place during sporulation shortly after σ^E activation and several genes under σ^E control are known to be important for cortex formation [13].

3.6.2. Lowest variance

We use the 433 genes with lowest variance (<0.03) of the posterior distribution of log-expression ratio in the mutant design as the subset of interest. This set includes genes both up- and down-regulated, and represents 38 functional subcategories. The population for this and the remaining tests is the set of 3712 genes detected on at least one array of the mutant design. Two subcategories were over-represented, with significant adjusted *p*-values (less than 0.01), including ''membrane bioenergetics'' and ''RNA modification''. Two more subcategories were over-represented, with less

significant p-values (<0.10) but still of interest, including "cell wall" and "metabolism of amino acids and related molecules" (Table 2). These results suggest that these are vital functions with strong expression control that does not vary across slides.

3.6.3. Highest variance

We use the 473 genes with highest variance (>0.25) in the posterior distribution of the log-expression ratios in the mutant design as the subset of interest. This subset includes both up- and down-regulated genes, and represents 37 functional subcategories. One subcategory was over-represented, with significant adjusted p-value (less than 0.01), which was the "unknown" functional category. This suggests that this group of genes may not be well studied due to their high variance and other characteristics that may not be reliably measured. However, it is difficult to draw firm conclusions, since the "unknown" functional category is by definition very heterogeneous. Most of these genes will be reassigned to existing categories in the future as research progresses.

3.6.4. Non-differentially expressed genes

We examined the 311 genes that were non-differentially expressed in the mutant design. These genes had scores between 0.45 and 0.55, and a range of posterior median log-expression ratios between $(-0.06, 0.06)$. This group represents 31 functional subcategories. We did not find any of the subcategories to be over-represented in this group of genes.

4. Discussion

We demonstrated here the usefulness of the Bayesian hierarchical model developed by Tseng et al. [14] in determining differentially expressed genes in a large-scale biological investigation, which involves designs with multiple slides in repeated experiments and a substantial amount of missing data.

There were genes that were overexpressed in only the mutant experiments and are of interest. The mutant experiments were easier to interpret than the induction experiments since the condition of sporulation used in the mutant design corresponds to the condition where σ^E is normally active. The genes of the induction experiments had lower expression ratios due to a more artificial design, i.e. σ^E is normally not expressed in growing cells. For this reason, there were approximately 50 genes that were overexpressed in the mutant design but not the induction design that have a high chance of being sporulation genes. However, the scope of this project was to identify targets of which the biologist was most certain.

The nearest neighbor scan-statistic used for the identification of chromosomal clusters of σ^E-controlled genes can be generally applied to find interesting chromosomal areas where a cluster of genes with similar characteristics are present. Hypotheses exist to explain our findings of higher density of σ^E-controlled genes in

certain regions of the bacterial chromosome. Part of the observed clustering is due to the fact that about half of the genes in the σ^E regulon are organized in operons. This form of clustering is very frequent in bacterial genomes and is therefore expected in this particular case as well. However, more surprising is the observation that not all of the observed clustering can be explained by the presence of operons. A possible explanation for this additional clustering is that most of the sporulation genes appeared at the same time in evolutionary history and therefore were clustered in the same region of the chromosome of the ancestral endospore former. The fact that these genes are still somewhat clustered could be a remnant of this ancient compact organization [6]. These questions can be further studied through comparative genomics, using closely related species that also sporulate, such as *Bacillus halodurans*, *Oceanobacillus iheyensis* and *Bacillus anthracis*, closely related non-endospore formers *Listeria monocytogenes* and *Listeria innocua*, and more distantly related endospore-formers *Clostridium acetobutylicum* and *Clostridium perfringens* (see [6] and references therein).

The large segment of the chromosome of 218 genes with no σ^E-controlled genes corresponds to the SPβ prophage. This prophage is present in the genome sequence of strain 168 [9], which is the parental strain that was used for genome sequencing and building of the microarrays. However, the SPβ prophage is absent from strain PY79 [16], a derivative of strain 168 that was used in transcriptional profiling experiments. Therefore, the spots corresponding to SPβ prophage on the microarray have been consistently flagged and removed from analysis. This results in no σ^E-controlled genes being found in this segment of the chromosome.

Acknowledgments

We thank George Tseng for providing R subroutines and helpful advice, Cristian Castillo-Davis for providing *B. subtilis* GeneMerge files and helpful direction, and Shane Jensen and Hosung Kang for helpful discussion. We are grateful to Richard Losick in whose laboratory the work of PE was carried out. We also thank two anonymous referees for helpful comments. This research was supported by the National Institutes of Health Grants 1F37LM07626-01 to EMC, and NSF Grant DMS-0204674 to JSL. PE was supported by a post-doctoral fellowship from the Swiss National Science Foundation and the Human Frontier Science Program and is currently supported by a Merck Core Educational Support Program. Research in Richard Losick's laboratory is supported by NIH Grant GM18568.

References

[1] B. Beall, C.P. Moran Jr., Cloning and characterization of spoVR, a gene from *Bacillus subtilis* involved in spore cortex formation, J. Bacteriol. 176 (1994) 2003–2012.

[2] C.I. Castillo-Davis, D.L. Hartl, GeneMerge—post-genomic analysis, data mining, and hypothesis testing, Bioinformatics 19 (2003) 891–892.

[3] E.M. Conlon, X.S. Liu, J.D. Lieb, J.S. Liu, Integrating regulatory motif discovery and genome-wide expression analysis, Proc. Nat. Acad. Sci. USA 100 (2003) 3339–3344.
[4] N. Cressie, On some properties of the scan statistic on the circle and the line, J. Appl. Probab. 14 (1977) 272–283.
[5] S. Dudoit, Y.H. Yang, M.J. Callow, T. Speed, Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, Statist. Sinica 12 (2002) 111–140.
[6] P. Eichenberger, S.T. Jensen, E.M. Conlon, C. van Ooij, J. Silvaggi, J.-E. Gonzalez-Pastor, M. Fujita, S. Ben-Yehuda, P. Stragier, J.S. Liu, R. Losick, The σ^{E} regulon and the identification of additional sporulation genes in *Bacillus subtilis*, J. Molecular Biol. 327 (2003) 945–972.
[7] J. Glaz, Approximations for the tail probabilities and moments of the scan statistic, Statist. Med. 12 (1993) 1853–1865.
[8] R. Ihaka, R. Gentleman, A language for data analysis and graphics, J. Comput. Graphical Statist. 5 (1996) 299–314.
[9] F. Kunst, et al., The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*, Nature 390 (1997) 249–256.
[10] J.S. Liu, Monte Carlo Strategies in Scientific Computing, Springer, New York, 2001.
[11] N.D. Neff, J.I. Naus, The Distribution of the Size of the Maximum Cluster of Points on a Line. Selected Tables in Mathematical Statistics, vol. VI, AMS, Providence, RI, 1980.
[12] E.E. Schadt, C. Li, B. Ellis, W.H. Wong, Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data, J. Cell Biochem. Suppl. 37 (2001) 120–125.
[13] P. Stragier, R. Losick, Molecular genetics of sporulation in *Bacillus subtilis*, Ann. Rev. Genetics 30 (1996) 297–341.
[14] G.C. Tseng, M.K. Oh, L. Rohlin, J.C. Liao, W.H. Wong, Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects, Nucleic Acids Res. 29 (2001) 2549–2557.
[15] Y.H. Yang, S. Dudoit, P. Luu, T.P. Speed, Normalization for cDNA microarray data, in: M.L. Bittner, Y. Chen, A.N. Dorsel, E.R. Dougherty (Eds.), Microarrays: Optical Technologies and Informatics, Proceedings of SPIE, Vol. 4266, SPIE, San Jose, CA, 2001, pp. 141–152.
[16] P. Youngman, J.B. Perkins, R. Losick, Construction of a cloning site near one end of Tn917 into which foreign DNA may be inserted without affecting transposition in *Bacillus subtilis* or expression of the transposon-borne *erm* gene, Plasmid 12 (1984) 1–9.
[17] Y. Zheng, J.D. Szustakowski, L. Fortnow, R.J. Roberts, S. Kasif, Computational identification of operons in microbial genomes, Genome Res. 12 (2002) 1221–1230.